Annotation Guideline for Single Nucleotide Polymorphisms

Philippe Thomas

July 17, 2009

This guideline aims to provide sufficient information to annotate articles describing Single Nucleotide Polymorphisms (SNP) and associate them to dbSNP identifier. More precisely only SNP substitutions are annotated. A SNP mentions encompasses *location* and two *alleles* (wildtype and mutated). In difference to other public available SNP corpora, protein sequence mutations and nucleic sequence mutations are annotated. Annotation is performed using WordFreak `http://wordfreak.sourceforge.net/` and SNP-span encompasses location and both alleles. Texts are automatically pre-annotated by MutationFinder, but all extracted mentions should be verified. Mentions missed by MutationFinder have to tagged manually. SNP mentions are also associated with the dbSNP identifier mentioned in the respective article. SNPs mentioned only in terms of a identifier (like rs334) or by textual description are discarded.

1 Location

Location specifies the exact location of a SNP on a gene or on the noncoding genomic background. In many publications SNPs are located on a speficic gene, however the gene mention itself (like BRCA1) is not part of the *location*.

1.1 Signs

Positive sign are part of the *location*, like: "at position +123". On negative signs its often harder to distinguish whether the sign is part of the location or used as a seperator to the preceding word. In these cases one has to read the text carfully. In the example *"The Ile-165 mutation did not confer..."* someone can easyly recognize that the text is describing a SNP located on a amino acid, as SNPs on amino acids can never have negative signs. However in the example *"We found a mutation on CYPD2-12 A→G"* it can not deduced wether the minus sign is part of the location or a seperator between the gene name and the position of the SNP. In ambigous cases, the minus sign is not part of the location.

1.2 Intervening Sequence

Mutations located on introns are sometimes described using the abbreviation IVS followed by the number of the intron and the distance to the closest exon. Typical mentions are *"The following mutations were not detected IVS 12+1 A>T and IVS 13-12 G>T."*

1.3 Typical Examples for Location

- ... found at amino acid 12 ...
- ... at IVS1+12 ...

2 Allele

Typically a SNP has at least two known states (a wildtype– and a mutated–allele).

2.1 Typical examples

Valid alleles are all nucleotides in one letter code or fullname mentions and all amino acids in one letter–, three letter–, triplet–code or fullname mentions. One exception is the stop codon where the valid abbreviation is the letter *X*. In some older publications the stop codon is sometimes refered as amber, ochre or opal codon.

- ... substitution from Leu to Pro
- We induced a opal codon at ...